Published for SISSA by Springer

Received: June 7, 2019
Revised: July 20, 2019
Accepted: September 20, 2019
Published: October 4, 2019

Adversarially-trained autoencoders for robust unsupervised new physics searches

Andrew Blance,[a,b] **Michael Spannowsky**[a] **and Philip Waite**[a]

[a] *Institute for Particle Physics Phenomenology, Department of Physics, Durham University, Durham, DH1 3LE, U.K.*

[b] *Institute for Data Science, Durham University, Durham, DH1 3LE, U.K.*

E-mail: andrew.t.blance@durham.ac.uk, michael.spannowsky@durham.ac.uk, p.a.waite@durham.ac.uk

Abstract: Machine learning techniques in particle physics are most powerful when they are trained directly on data, to avoid sensitivity to theoretical uncertainties or an underlying bias on the expected signal. To be able to train on data in searches for new physics, anomaly detection methods are imperative, which can be realised by an autoencoder acting as an unsupervised classifier. The last source of uncertainties affecting the classifier are then experimental uncertainties in the reconstruction of the final-state objects. To mitigate their effect on the classifier and to allow for a realistic assessment of the method, we propose to combine the autoencoder with an adversarial neural network to remove its sensitivity to the smearing of the final-state objects. We quantify its effect and show that one can achieve a robust anomaly detection in resonance-induced $t\bar{t}$ final states.

Keywords: Beyond Standard Model, Particle correlations and fluctuations, Jet physics, Top physics, Hadron-Hadron scattering (experiments)

ArXiv ePrint: 1905.10384

https://doi.org/10.1007/JHEP10(2019)047

JHEP10(2019)047

Contents

JHEP10(2019)047

1 Introduction

In recent years machine learning algorithms, and in particular neural networks, have become increasingly popular in analysing large quantities of data. In the context of particle physics two main applications are prevalent: the classification of data according to different hypotheses [1–33] and the regression of data to interpolate and extrapolate object-relevant properties [34–38].

Using multi-variate analysis (MVA) techniques to classify events into signal and background classes based on their radiation profiles should improve the LHC's experiments' sensitivity in searches for new physics. Machine learning algorithms are able to analyse multiple observables or inputs simultaneously to find a region in this multi-dimensional parameter space that shows a relative enhancement of signal over background events. To find this region in a supervised-learning approach, pseudo-data for signal and background need to be generated using event generators, e.g. Sherpa [39], Herwig [40] or Pythia [41], and the respective training samples are made known to the algorithm whether they contain signal or background respectively. However, as the Monte Carlo event samples are plagued by theoretical uncertainties, the classification algorithm will be subjected to the same uncertainties. This issue is amplified by the fact that the MVA method will usually favour highly-exclusive phase space regions which are poorly understood perturbatively [42, 43], and often observables that are not even IR-safe are found in experimental measurements to be most discriminative, e.g. the number of charged tracks [44, 45], thus further questioning the reliability of theoretically predicted classification efficiencies. Adversarial neural networks have been proposed to desensitise classification methods against theoretical [46] and systematic uncertainties [47] or against certain observables [48].

One promising approach to overcome deficits from training on pseudo-data is to train on actual data directly.[1] While so-called data-driven methods are not subjected to theoretical uncertainties, one has to make sure that signal and background are sufficiently pure to train the algorithm on well-separated event samples. Most of the time, and in particular in searches for new physics, this is a highly challenging task. Rare processes, e.g. the production of di-Higgs final states, or completely unknown processes, e.g. the production of a gluino, are of utmost interest to search for at the LHC. However, obtaining a data-driven training sample for such processes is impossible, thus, limiting the applicability for data-driven methods. One way around this bottleneck is not to train on signal at all, but to identify the kinematic features of background samples and to design a method that flags up events that do not possess the same features, thereby classifying such an event as signal. The remaining residual experimental problem that remains for such an approach are the experimental and systematic uncertainties in the measurements of the inputs of a data-driven anomaly detection method.

Autoencoders [60, 61] have been proposed for denoising [62], generative models [63] and in particular for anomaly detection [48, 64–66]. They use an information bottleneck to map an input to a latent-compressed representation and then decode this representation back. The loss function measures the squared difference between input and decoded output. By minimising the loss function, the autoencoder learns intrinsic features of the training samples that survive the information bottleneck. After training the autoencoder on background samples, it is expected that applying the autoencoder to signal samples will result in a modified value for the loss function, as some kinematic features differ between signal and background. The depth of the networks and the width of the bottleneck are hyperparameters of the network that can be optimised for the problem at hand. Using autoencoders for anomaly detection, we will show that adversarially-trained neural networks can take systematic uncertainties into account and desensitise the classification performance in data-driven searches for new physics. To achieve this, we adversarially train an autoencoder on Monte-Carlo-generated pseudo-data which has been systematically smeared in order for it to learn to reconstruct the events without using any information about the smearing.

We apply this framework to resonance searches, i.e. a heavy colour-singlet scalar, colour-octet scalar and colour-singlet vector, that are well-motivated by many new physics models. This selection allows one to study the impact of the spin and colour quantum numbers of the resonances on the classification efficiencies.[2] The resonances are assumed to subsequently decay into top quarks [68–71]. Top quark samples are an ideal playground for anomaly detection, as they can be purified to a very high degree, i.e. the confidence that one trains on a pure $t\bar{t}$ sample is very high, in particular when one top decays hadronically while the other decays leptonically. On the other hand, top final states are complex, consisting of many jets, leptons and missing transverse energy. Thus, uncertainties on reconstructed observables due to detector effects can be large.

[1] If machine learning techniques can be trained on data directly they become independent of theoretical uncertainties. In such circumstances they can outperform theory-based reconstruction approaches, like the matrix element method [49–53], which was recently extended to fully exclusive final states [54–59].

[2] The quantum numbers of the decaying resonances are known to have a strong impact on the reconstruction efficiencies of boosted top quarks [67].

JHEP10(2019)047

The paper is structured as follows: in section 2 we first discuss the analysis setup. To establish a baseline of what can be achieved by supervised learning, we show the performance of a neural network classifier and the effect of combining it with an adversarial neural network in section 3. In section 4 we extend this approach to an unsupervised autoencoder for anomaly detection and consider its application to other new physics models and the effects of an impure training sample. We offer conclusions on our findings in section 5.

2 Analysis setup and smearing procedure

We use MadGraph5_aMC@NLO [72] to generate the events for the study, followed by Pythia 8.2 for parton shower and hadronisation. The background events consist of $pp \to t\bar{t}$ at a centre-of-mass energy of 14 TeV, with one top quark forced to decay leptonically and the other hadronically. The signal events are generated from a heavy Z' boson [73] via $pp \to Z' \to t\bar{t}$, also with semileptonic decays of the top quarks. As a benchmark for this study, we select the Z' mass to be 2 TeV with a width of 89.6 GeV. A transverse momentum cut of $p_T > 500$ GeV is applied directly to the top quarks at generator level, for both signal and background events.

Following the concept of reconstructing highly boosted top quarks with fat jets [74, 75], the hadrons and non-isolated leptons from the event are initially clustered into jets using the Cambridge-Aachen algorithm [76] with a radius of $R = 1.0$. The constituents of the two hardest fat jets are then reclustered into jets using the k_T algorithm with $R = 0.2$, implemented in FastJet [77]. Jets are required to have $p_T > 30$ GeV and are b-tagged through their association to a B-meson. Isolated leptons are required to have $p_T > 10$ GeV. Events are selected which have a scalar-summed visible transverse momentum of $H_T > 1$ TeV, and which have at least one b-jet inside one fat jet, at least one b-jet and two light jets inside the other fat jet, and at least one isolated lepton.

The observables that we consider for the analysis are the four-momenta of the two b-jets, two light jets and isolated lepton, as well as the missing energy ($\not{E}_T$) in the event (21 observables in total). To represent possible systematic uncertainties that can arise in detectors from jet energy scales, we apply a smearing procedure to the jets and the missing energy in the events. For the jets and leptons, we use a smearing based on refs. [78, 79] where the three-momenta of each object is smeared with a Gaussian. In our case, we take the extremities of this Gaussian so that the smearing is either applied upwards or downwards for all objects, with the relative width of the smearing envelope being larger for smaller p_T values. Similarly, we apply a shift to the missing energy based on ref. [80], where the width of the shift is proportional to $\sqrt{H_T}$, and use the two extremities of the envelope. We fix the direction of the missing energy smearing to always be the same as that of the jets and leptons. For the purposes of this study, we increase the size of the smearing envelope by a further factor of three, to be conservative on the systematic uncertainties and highlight the ability of our setup to correct for it.

We apply the smearing to the background sample such that two extra datasets are created for smearing in the upwards and downwards directions, as well as the unsmeared central sample. No smearing is applied to the signal sample. The three background samples

JHEP10(2019)047

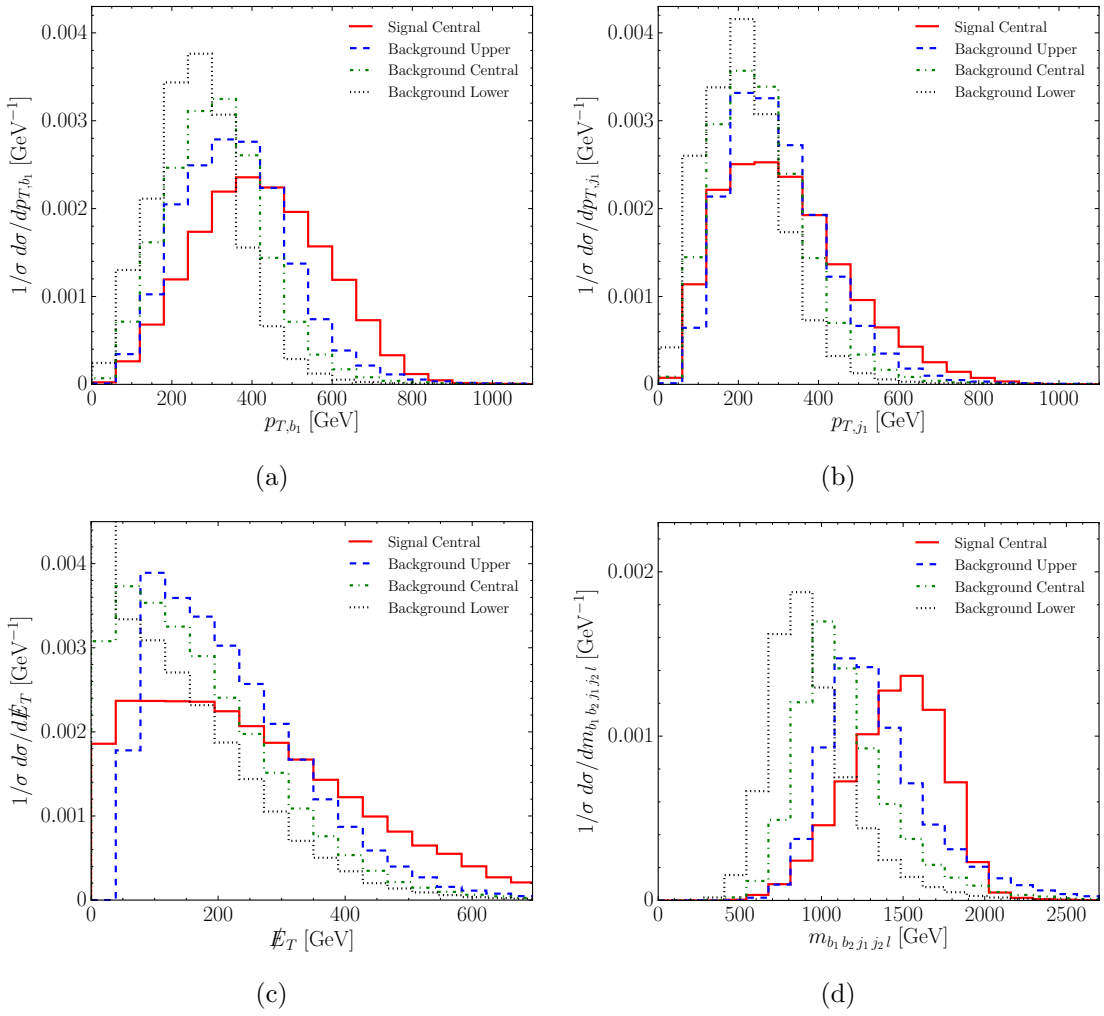


Figure 1. Effect of smearing on (a) the p_T of the hardest b-jet, (b) the p_T of the hardest light jet, (c) the missing energy and (d) the invariant mass of the jets and lepton, compared to the unsmeared background and the signal samples.

are each generated from statistically independent generator samples, and after all cuts we select 100,000 events from each of the four samples, with 20% of these retained for testing.

In figure 1 we show the effect of smearing on the p_T of the hardest b-jet and light jet, the missing energy in the event and the invariant mass of the jets and lepton, compared to the equivalent distributions for the signal events. Clearly the smearing of the background has the potential to make it either easier or harder for a classifier to discriminate between signal and background, depending on which direction the smearing shifts the background distribution.

3 Decorrelated jet smearing with supervised adversarial classifier

To set a benchmark for the signal-to-background separation, we first train a simple neural network classifier to discriminate signal events from the complete set of background events

JHEP10(2019)047

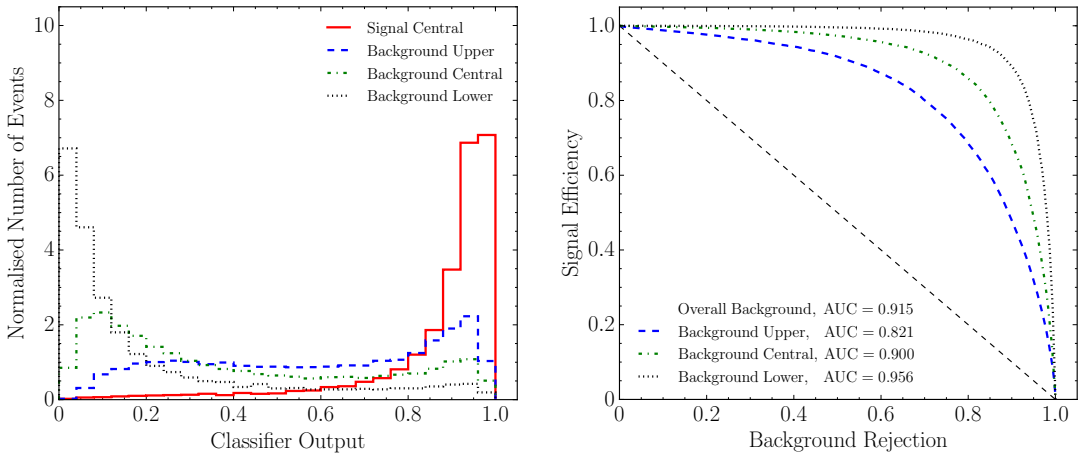


Figure 2. Supervised neural network classifier output (a) and ROC curves (b) for a classifier trained to classify signal and background events. The three background distributions result from the three different directions of smearing.

(including all three samples). We expect this supervised-learning approach to perform better than the unsupervised approach which follows.

The network consists of two hidden layers each with 20 nodes, with ReLu activations, and a final layer with a single sigmoid output. We use a binary cross entropy loss function since there are two possible classes. A class weighting in the loss function is used to account for the higher frequency of background events in the training data, i.e. the loss of the signal events are weighted higher. The network is trained using the Adam optimiser [81] with a learning rate of 0.01 and a batch size of 500 for 500 epochs. The network is implemented in Keras [82] with a TensorFlow [83] backend, and we use these throughout the rest of this paper. The results are shown by the distributions of the classifier outputs and the corresponding receiver operating characteristic (ROC) curves in figure 2. These are obtained by testing the network on each of the three background sets separately, and performing a classification against the central signal sample for each one. Also shown are the area-under-curve (AUC) scores for each curve as well as the score for all the background test samples combined. The network performance is strongly dependent on the direction that the sample has been smeared in. This can be understood from the observables in figure 1 where there is a larger overlap between the signal distribution and the background which has been smeared upwards.

We now extend this classifier with an adversarial network which is designed to discriminate the smearing class that the background sample came from, based upon the output of the classifier. The aim for such an extension to the classifier is to attempt to remove such a large dependence of its performance on the smearing of the background [46, 47]. The adversary and classifier are forced to take part in a zero-sum game — the classifier must learn to make its prediction without using any information derived from the smearing, in order to make it as hard as possible for the adversary to be able to discriminate the

JHEP10(2019)047

background samples. This is achieved by the two networks having opposite optimisation objectives, so that the classifier is penalised when the adversary performs better.

The adversarial network consists of two hidden layers with 20 nodes and ReLu activation functions, and takes as an input the output of the classifier. The output of the adversary has three nodes (one for each smearing class) with a softmax activation function and a categorical cross entropy loss. The network is then trained as follows:

1. The classifier is trained for three epochs using the Adam optimiser with a learning rate of 0.01 and a batch size of 500. A class weighting is applied to account for the higher frequency of background events in the training data.

2. The adversary is trained on background events for three epochs using mini-batch gradient descent with a learning rate of 0.01 and a batch size of 500.

3. The classifier is trained for one epoch with mini-batch gradient descent with a batch size of 500 and with a total loss function,

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{class}} - \alpha \mathcal{L}_{\text{adv}} \ . \tag{3.1}$$

   Furthermore, two class weightings are applied: one to account for the higher frequency of background events that the classifier is trained on, and one to account for the fact that the signal events are unsmeared, resulting in a higher frequency of unsmeared events that the adversary is trained on.

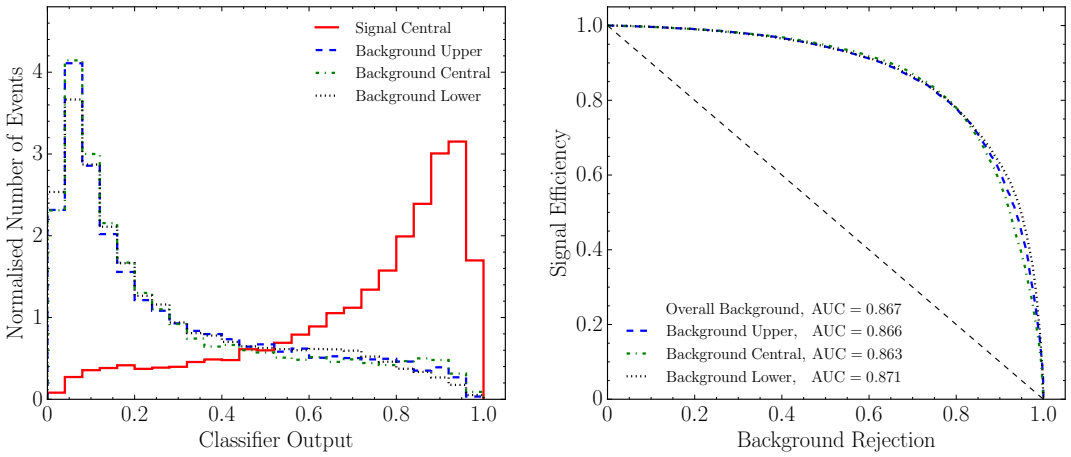
4. The adversary is trained on background events for one epoch using mini-batch gradient descent with a batch size of 500.

5. Steps 3 and 4 are repeated until they have been performed a total of 1000 times, with the learning rate decaying every 100 epochs to a factor of 0.75 of its previous value, starting from an initial value of 0.01.

The weight factor α in eq. (3.1) determines the relative importance of the two optimisation objectives. If it is set to zero, then the adversary has no effect on the training of the classifier. If it is too large, however, the performance of the classifier is severely affected. We find a value of 100 works well for our setup. There is another approach to training the adversarial network, where one updates the weights of both networks simultaneously. However, we find the approach of alternating the training — where the classifier is trained with the adversary weights frozen, and vice versa — to be more stable.

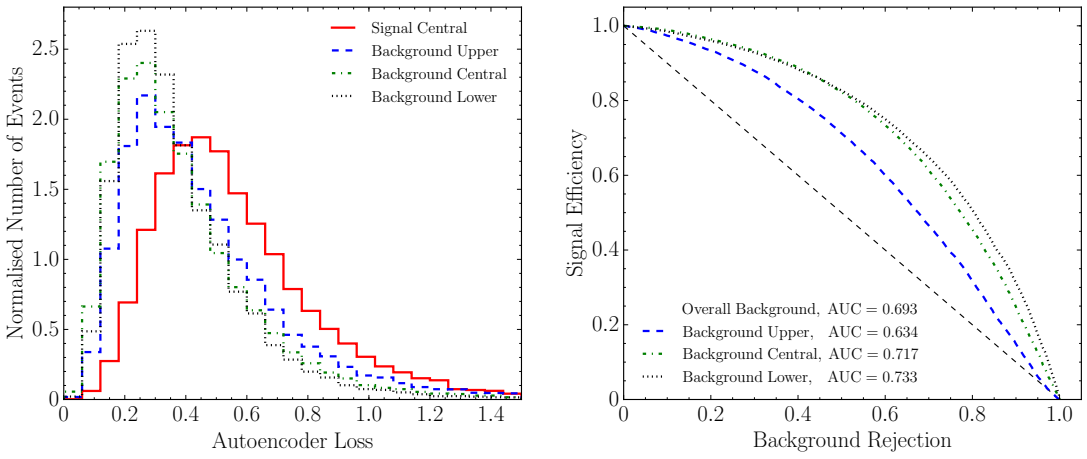
In figure 3, we show the performance of the adversarial classifier through the classifier output and ROC curves. The adversary has clearly had the effect of shaping the classifier outputs such that their dependence on the background smearing has been almost entirely removed. Thus, the ROC curves and AUC scores become very close together since the classification performance is now barely affected by the smearing.

JHEP10(2019)047



(a) (b)

Figure 3. Supervised neural network classifier output (a) and ROC curves (b) for an adversarial classifier trained to classify signal and background events. The three background distributions result from the three different directions of smearing.



(a) (b)

Figure 4. Autoencoder loss (a) and ROC curves (b) for an autoencoder trained only on background events. The three background distributions result from the three different directions of smearing.

4 Extension to unsupervised autoencoder

4.1 Adversarial autoencoder

As described earlier, autoencoders are an unsupervised learning algorithm which can be used as anomaly detectors to search for new physics since they only need to be trained on the background.

To this aim, we consider an autoencoder constructed from three hidden layers with 10, 3 and 10 nodes respectively, each with sigmoid activation functions. After the hidden layers, there is a linear output layer with the same dimension as the number of inputs, which

JHEP10(2019)047

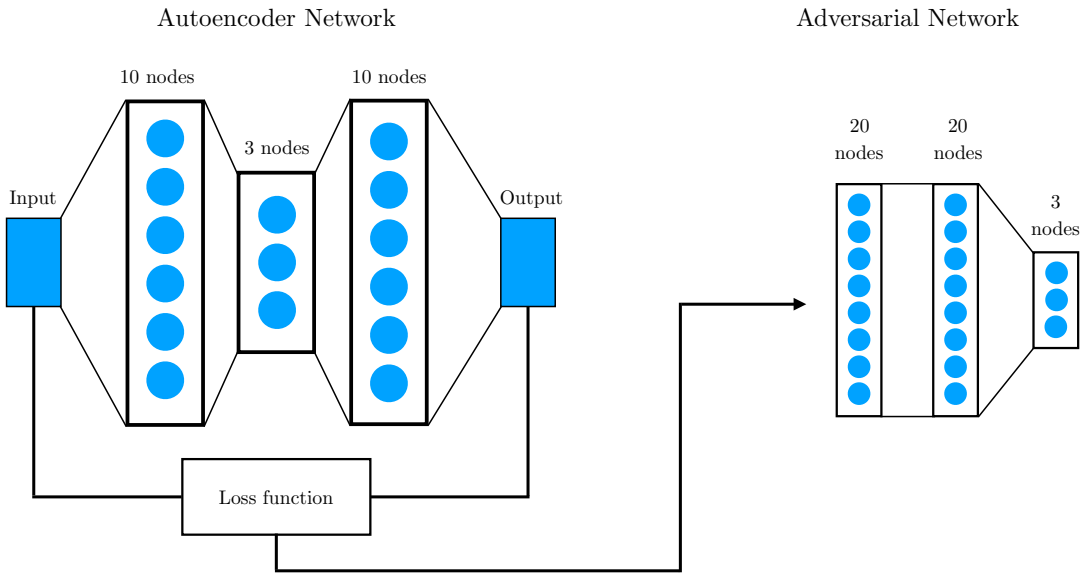


Figure 5. Architecture of the adversarial autoencoder. The loss function of the autoencoder is used as an input to the adversary for it to discriminate the smeared background samples.

correspond to the 21 observables. The loss is the mean squared error between the inputs and outputs — namely, the autoencoder has the goal of reconstructing the inputs as well as possible, having encoded the information into the latent-compressed layer. We train the autoencoder on the three background samples using the Adam optimiser with a learning rate of 0.01 for 500 epochs, and the results are shown in figure 4. Since the autoencoder is trained only on the background events, it learns how to reconstruct background events better than the signal events, and so the distribution of the losses for the signal events in figure 4(a) is at higher values. The ROC curves in figure 4(b) are obtained by performing a cut on the loss function and labelling all events above the cut as signal events, and all events below the cut as background events, and then moving this threshold across all values. This is similar to how the ROC curves are calculated from the output of the classifier, where the threshold is varied between 0 and 1 instead.

As we saw for the classifier, the smearing of the background has an effect on how well the autoencoder can be used to classify events, with the events which have been smeared upwards being mislabelled as signal events more often. It is important to note that the overall classification performance of the autoencoder is much worse than for the dedicated supervised classifier in section 3. However, this is not surprising — the autoencoder is only ever trained on background events, and only sees the signal events during testing. Thus, for the separation between signal and background it learns the intricate kinematic features of the background only. Furthermore, the optimisation objective of the classifier is for it to achieve a strong classification performance, which is not the case for the autoencoder.

We now combine the autoencoder with an adversarial network to improve the reliability and robustness of this unsupervised-learning approach. To achieve the aim of the autoencoder being able to make its predictions independent of the smearing of the back-

JHEP10(2019)047

ground, we use the autoencoder loss as an input to the adversary. Since a threshold on the autoencoder loss is used to perform the classification between signal and background, it is completely analogous to the output of the dedicated classifier used above, on which a cut is placed to classify the events. This input is then followed by two hidden layers each with 20 nodes and ReLu activation functions, with three softmax output nodes and a categorical cross entropy loss. This architecture is illustrated by the diagram in figure 5. The training proceeds similarly to the adversarial classifier, but with only background events in the training sample:

1. The autoencoder is trained for three epochs using the Adam optimiser with a learning rate of 0.01 and a batch size of 500.

2. The adversary is trained for three epochs using mini-batch gradient descent with a learning rate of 0.01 and a batch size of 500.

3. The autoencoder is trained for one epoch with mini-batch gradient descent with a batch size of 500 and with a total loss function,

$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{auto}} - \alpha \mathcal{L}_{\text{adv}} \ . \tag{4.1}$$

4. The adversary is trained for one epoch using mini-batch gradient descent with a batch size of 500.

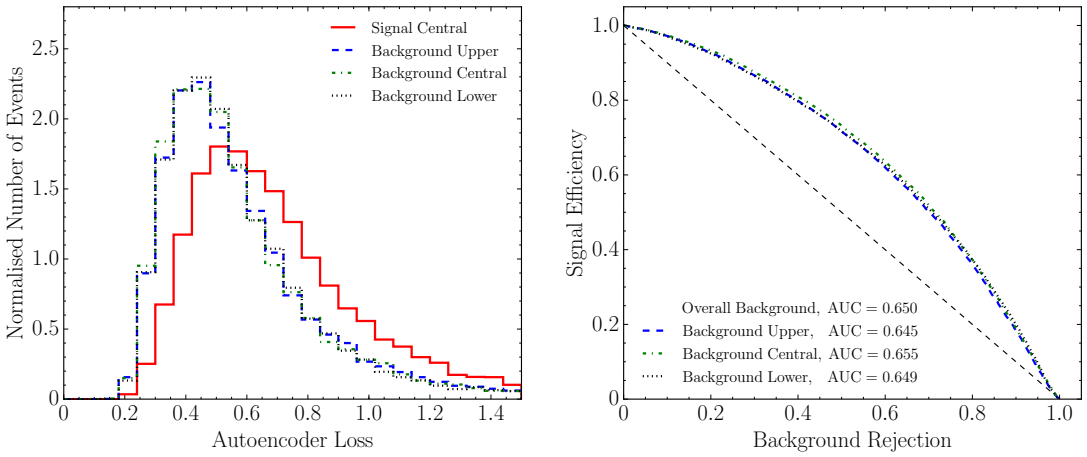
5. Steps 3 and 4 are repeated until they have been performed a total of 1500 times, with the learning rate decaying every 100 epochs to a factor of 0.75 of its previous value, starting from an initial value of 0.01.

We find this procedure to provide stable and numerically reliable results. Again, the relative weighting between the autoencoder and the adversary is set to $\alpha = 100$. The performance of the adversarially-trained autoencoder is shown in figure 6. The background distributions shown in figure 6(a) have been shaped such that they are independent of the direction of smearing, which results in the ROC curves in figure 6(b) becoming almost identical. This shows that the method has become independent of uncertainties inherent to the reconstruction of the final-state objects of LHC events.

In addition, we note that our setup also has the ability to interpolate to smaller amounts of smearing — although we have trained using background data which has been systematically smeared by a very large amount, we find that if it is tested on samples which have been smeared by a much smaller amount (without the increase by a factor of three), then the output of the adversarially-trained autoencoder (and also for the classifier in the previous section) is still insensitive to the smearing. Furthermore, we find that the AUC score increases when it is tested on a smaller amount of smearing, and is similar to the result of having both trained and tested it on this smaller amount. Therefore, the fact that the adversary is trained on a larger amount of smearing than is realistic does not adversely affect its performance.

We will now briefly recap what we have achieved by combining an autoencoder with an adversarial neural network. We started with three sets of background events — one

JHEP10(2019)047



(a) (b)

Figure 6. Autoencoder loss (a) and ROC curves (b) for an adversarial autoencoder trained only on background events. The three background distributions result from the three different directions of smearing.

which had been smeared upwards, one which had been smeared downwards, and one which had not been smeared at all. This smearing corresponded to the extremities of a Gaussian envelope, and was applied to jets, leptons and the missing energy in each event accordingly. Furthermore, we also had a set of signal events which had not been smeared. The smearing had the effect of shifting the kinematic features of the background such that the events which had been smeared upwards looked more like signal events, and the ones which had been smeared downwards looked less like signal events. This can be seen from the distributions in figure 1.

We then trained an autoencoder on all the background events for the purpose of using it to detect signal events, which have a higher expected reconstruction loss. In figure 4(b), the ROC curves are the result of testing the classification performance of the autoencoder for the signal separately against each background, and as expected, the autoencoder had a harder time discriminating the signal events against background events which had been smeared upwards. We then combined this with an adversarial neural network, which had the objective of recognising which direction each background sample had been smeared in based upon the loss of the autoencoder. The autoencoder and adversary were trained using a combined loss function, which penalised the autoencoder for outputting reconstruction losses from which the adversary could discriminate the samples. The result of this is that the autoencoder has learnt to reconstruct events without using any information derived from the smearing, which can be seen from the fact that the ROC curves in figure 6(b) have converged.

4.2 Corrupted autoencoder and application to other new physics models

Thus far, the analysis has been carried out on training sets consisting of pure background events. Realistically, data may not actually look like this since if new physics exists, then

JHEP10(2019)047

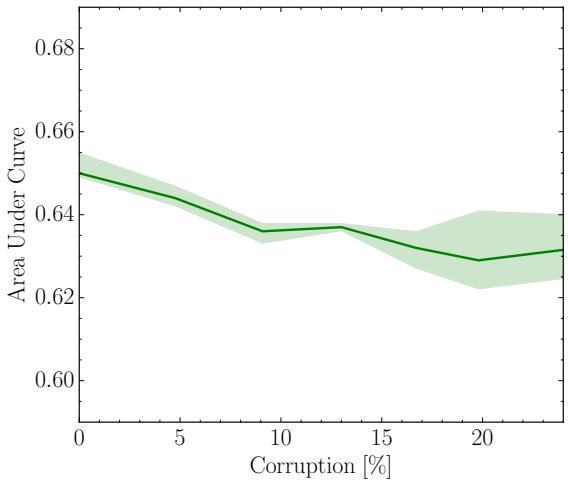


Figure 7. Effect of contaminating the training sample with an increasing fraction of signal events. The central line shows the overall AUC score, and the band represents the difference between the upper and lower AUC scores.

it would also form part of that same data. To begin accounting for this it is possible to inject into the three background sets appropriately smeared signal events. By training on these newly contaminated sets we can investigate how sensitive the performance of the adversarial autoencoder is to an increase in signal corruption in the training set. In figure 7, we show these results. The band represents the difference between the upper and lower AUC scores, which shows how well the adversary desensitises the autoencoder from the smearing, and the central line is the overall AUC score. All model hyperparameters are left unchanged during the training, with only the relative fraction of corruption changing, defined as a percentage of the total training set. From the plot it is clear that injecting signal events during training has little effect on the overall performance until the fraction of corruption becomes unrealistically large, showing the potential applicability of the method to real data.

Since the performance is not drastically affected by a corruption of the training data, we can proceed with a training sample consisting purely of background events. One of the advantages of the autoencoder only needing to be trained on background events is that it can then be tested for signal events arising from any model. Here, we test our adversarially-trained autoencoder on a variety of different new physics models. We aim to quantify the effect of the resonance's quantum numbers, i.e. spin, colour and coupling strengths, on the performance of the autoencoder. The models used are:

- Two further Z' cases with widths of 10 GeV and 200 GeV. In both cases the masses are held at 2 TeV.
- A scalar colour-octet [84], with a mass of 2 TeV and the scalar and axial parameters fixed to ensure the width is ~ 89.6 GeV.
- A scalar colour-singlet with a mass of 2 TeV and a width of 89.6 GeV.

JHEP10(2019)047

Signal	Overall AUC	Upper-Lower Difference	Cross Section Limit [pb]
$Z'_{w=10\ \text{GeV}}$	0.662	0.009	0.0101
$Z'_{w=89.6\ \text{GeV}}$	0.656	0.009	0.0098
$Z'_{w=200\ \text{GeV}}$	0.650	0.009	0.0105
Scalar	0.654	0.010	0.0104
Octet	0.659	0.010	0.0102

Table 1. The overall AUC score, difference between the largest and smallest AUC scores and the cross section limits found from using the adversarial autoencoder trained only on background events and tested on the original Z' case and four other signals.

Table 1 shows the results of testing the adversarially-trained autoencoder on the new signals. In each case the adversary is able to perform well, with the difference between the upper and lower AUC scores showing that the new signals do not hinder the ability of adversary to desensitise the autoencoder to the smearing. This behaviour is of course expected, since the same background samples are used to test against each new signal. We also show estimates of the potential limits on the cross sections that can be obtained using the classification performance of the autoencoder. These are calculated by finding the points on the ROC curves that maximise $S/\sqrt{B}$, then comparing them to the background cross section and assuming an integrated luminosity of $100\,\text{fb}^{-1}$. We then require that $S/\sqrt{B} > 2$ to set a 95% confidence limit. The limits we find are insensitive to the nature of the resonance i.e. with respect to their quantum numbers, and they are comparable to the limits found by ATLAS in ref. [69].[3]

5 Conclusions

The ideal scenario for the usage of machine learning methods is when they can be applied directly on experimental data, without the requirement to train them on pseudo-data or without theoretically calculated inputs, e.g. as in the Matrix Element Method. In such circumstances neither theoretical uncertainties that challenge the robustness of the method, nor a theoretical bias regarding the features of the signal are introduced. Thus, the experimental data alone would be sufficient to identify anomalous events, which could be isolated and studied further to discover new physics. Such identification of anomalous events can be realised using an autoencoder, trained on a pure background sample. However, even in this ideal scenario, residual uncertainties due to the imperfect reconstruction of final-state objects remain.

Focusing on resonance searches in semileptonic $t\bar{t}$ final states, we quantified the performance of an adversarially-trained autoencoder. In particular, we compared the performance of an autoencoder-based unsupervised-learning approach with a supervised neural

[3]However, note that we show the new physics cross section after event selection and reconstruction cuts, while ATLAS shows the inclusive cross section for a specific Z' model. Furthermore, our analysis was performed at 14 TeV, while the limits from ATLAS have been obtained at a centre-of-mass energy of 13 TeV.

JHEP10(2019)047

network classifier. While the supervised classifier performs significantly better than the unsupervised-learning approach, the latter still shows a strong aptitude in telling apart signal from background events. In both cases reconstruction uncertainties show however a big impact on the evaluated performance of the classifiers, thereby evidencing the need for measures to desensitise them against such uncertainties for a reliable performance evaluation.

We proposed to combine the autoencoder with an adversarial neural network to realise a robust and reliable unsupervised anomaly detection method that can be readily applied to experimental data. The classification result is independent of the smearing of the reconstructed final-state objects over the entire range of the ROC curve and even extends to training on corrupted backgrounds, i.e. backgrounds with a large admixture of signal events. Although we applied it to Monte-Carlo-generated pseudo-data, we envisage that the procedure could be applied analogously to experimental data by creating labelled datasets that have been systematically smeared. Thus, this setup proves to be a very robust data-driven way to search for new physics resonances, irrespective of their quantum numbers, i.e. spin, colour or width.

Acknowledgments

MS acknowledges the generous hospitality of Barbara Jaeger and her group at the University of Tuebingen, as well as support of the Humboldt Society, during the completion of parts of this work.

References

[1] B. Nachman et al., *Jets from jets: re-clustering as a tool for large radius jet reconstruction and grooming at the LHC*, *JHEP* **02** (2015) 075 [arXiv:1407.2922] [INSPIRE].

[2] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[3] J. Barnard, E.N. Dawe, M.J. Dolan and N. Rajcic, *Parton shower uncertainties in jet substructure analyses with deep neural networks*, *Phys. Rev.* **D 95** (2017) 014018 [arXiv:1609.00607] [INSPIRE].

[4] L.M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly supervised classification in high energy physics*, *JHEP* **05** (2017) 145 [arXiv:1702.00414] [INSPIRE].

[5] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned top tagging with a Lorentz layer*, *SciPost Phys.* **5** (2018) 028 [arXiv:1707.08966] [INSPIRE].

[6] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) learning to do more with less*, *JHEP* **02** (2018) 034 [arXiv:1706.09451] [INSPIRE].

[7] S. Chang, T. Cohen and B. Ostdiek, *What is the machine learning?*, *Phys. Rev.* **D 97** (2018) 056009 [arXiv:1709.10106] [INSPIRE].

JHEP10(2019)047

[8] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet constituents for deep neural network based top quark tagging*, arXiv:1704.02124 [INSPIRE].

[9] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-aware recursive neural networks for jet physics*, *JHEP* **01** (2019) 057 [arXiv:1702.00748] [INSPIRE].

[10] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning top taggers or the end of QCD?*, *JHEP* **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[11] L. de Oliveira, M. Paganini and B. Nachman, *Learning particle physics by example: location-aware generative adversarial networks for physics synthesis*, *Comput. Softw. Big Sci.* **1** (2017) 4 [arXiv:1701.05927] [INSPIRE].

[12] H. Lüo, M.-x. Luo, K. Wang, T. Xu and G. Zhu, *Quark jet versus gluon jet: fully-connected neural networks with high-level features*, *Sci. China Phys. Mech. Astron.* **62** (2019) 991011 [arXiv:1712.03634] [INSPIRE].

[13] K. Datta and A.J. Larkoski, *Novel jet observables from machine learning*, *JHEP* **03** (2018) 086 [arXiv:1710.01305] [INSPIRE].

[14] A.J. Larkoski, I. Moult and B. Nachman, *Jet substructure at the large hadron collider: a review of recent advances in theory and machine learning*, arXiv:1709.04464 [INSPIRE].

[15] C. Shimmin et al., *Decorrelated jet substructure tagging using adversarial neural networks*, *Phys. Rev.* **D 96** (2017) 074034 [arXiv:1703.03507] [INSPIRE].

[16] E.M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, *JHEP* **10** (2017) 174 [arXiv:1708.02949] [INSPIRE].

[17] T. Roxlo and M. Reece, *Opening the black box of neural nets: case studies in stop/top discrimination*, arXiv:1804.09278 [INSPIRE].

[18] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *Constraining effective field theories with machine learning*, *Phys. Rev. Lett.* **121** (2018) 111801 [arXiv:1805.00013] [INSPIRE].

[19] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *A guide to constraining effective field theories with machine learning*, *Phys. Rev.* **D 98** (2018) 052004 [arXiv:1805.00020] [INSPIRE].

[20] J.H. Collins, K. Howe and B. Nachman, *Anomaly detection for resonant new physics with machine learning*, *Phys. Rev. Lett.* **121** (2018) 241803 [arXiv:1805.02664] [INSPIRE].

[21] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, 2018 *JINST* **13** P07027 [arXiv:1804.06913] [INSPIRE].

[22] K. Fraser and M.D. Schwartz, *Jet charge and machine learning*, *JHEP* **10** (2018) 093 [arXiv:1803.08066] [INSPIRE].

[23] P.T. Komiske, E.M. Metodiev, B. Nachman and M.D. Schwartz, *Learning to classify from impure samples with high-dimensional data*, *Phys. Rev.* **D 98** (2018) 011502 [arXiv:1801.10158] [INSPIRE].

[24] S. Macaluso and D. Shih, *Pulling out all the tops with computer vision and deep learning*, *JHEP* **10** (2018) 121 [arXiv:1803.00107] [INSPIRE].

[25] A. Andreassen, I. Feige, C. Frye and M.D. Schwartz, *JUNIPR: a framework for unsupervised machine learning in particle physics*, *Eur. Phys. J.* **C 79** (2019) 102 [arXiv:1804.09720] [INSPIRE].

JHEP10(2019)047

[26] P. De Castro and T. Dorigo, *INFERNO: inference-aware neural optimisation*, *Comput. Phys. Commun.* **244** (2019) 170 [arXiv:1806.04743] [INSPIRE].

[27] R.T. D'Agnolo and A. Wulzer, *Learning new physics from a machine*, *Phys. Rev.* **D 99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].

[28] J. Brehmer, G. Louppe, J. Pavez and K. Cranmer, *Mining gold from implicit models to improve likelihood-free inference*, arXiv:1805.12244 [INSPIRE].

[29] J.W. Monk, *Deep learning as a parton shower*, *JHEP* **12** (2018) 021 [arXiv:1807.03685] [INSPIRE].

[30] L. Moore, K. Nordström, S. Varma and M. Fairbairn, *Reports of my demise are greatly exaggerated: N-subjettiness taggers take on jet images*, arXiv:1807.04769 [INSPIRE].

[31] A. De Simone and T. Jacques, *Guiding new physics searches with unsupervised learning*, *Eur. Phys. J.* **C 79** (2019) 289 [arXiv:1807.06038] [INSPIRE].

[32] S. Bollweg et al., *Deep-learning jets with uncertainties and more*, arXiv:1904.10004 [INSPIRE].

[33] O. Cerri et al., *Variational autoencoders for new physics mining at the Large Hadron Collider*, *JHEP* **05** (2019) 036 [arXiv:1811.10276] [INSPIRE].

[34] ATLAS collaboration, *Generalized numerical inversion: a neural network approach to jet calibration*, ATL-PHYS-PUB-2018-013 (2018).

[35] ATLAS collaboration, *Performance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2*, *Eur. Phys. J.* **C 77** (2017) 673 [arXiv:1704.07983] [INSPIRE].

[36] CMS collaboration, *Performance of the CMS missing transverse momentum reconstruction in pp data at* $\sqrt{s} = 8$ *TeV*, 2015 *JINST* **10** P02006 [arXiv:1411.0511] [INSPIRE].

[37] CMS collaboration, *Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at* $\sqrt{s} = 8$ *TeV*, 2015 *JINST* **10** P06005 [arXiv:1502.02701] [INSPIRE].

[38] CMS collaboration, *Performance of photon reconstruction and identification with the cms detector in proton-proton collisions at* $\sqrt{s} = 8$ *TeV*, 2015 *JINST* **10** P08010 [arXiv:1502.02702] [INSPIRE].

[39] T. Gleisberg et al., *Event generation with SHERPA 1.1*, *JHEP* **02** (2009) 007 [arXiv:0811.4622] [INSPIRE].

[40] J. Bellm et al., *HERWIG 7.0/HERWIG++ 3.0 release note*, *Eur. Phys. J.* **C 76** (2016) 196 [arXiv:1512.01178] [INSPIRE].

[41] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[42] C. Englert, R. Kogler, H. Schulz and M. Spannowsky, *Higgs characterisation in the presence of theoretical uncertainties and invisible decays*, *Eur. Phys. J.* **C 77** (2017) 789 [arXiv:1708.06355] [INSPIRE].

[43] C. Englert, P. Galler, A. Pilkington and M. Spannowsky, *Approaching robust EFT limits for CP-violation in the Higgs sector*, *Phys. Rev.* **D 99** (2019) 095007 [arXiv:1901.05982] [INSPIRE].

[44] S. Schaetzel and M. Spannowsky, *Tagging highly boosted top quarks*, *Phys. Rev.* **D 89** (2014) 014007 [arXiv:1308.0540] [INSPIRE].

JHEP10(2019)047

[45] ATLAS collaboration, *Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *Eur. Phys. J.* **C 74** (2014) 3023 [arXiv:1405.6583] [INSPIRE].

[46] C. Englert, P. Galler, P. Harris and M. Spannowsky, *Machine learning uncertainties with adversarial neural networks*, *Eur. Phys. J.* **C 79** (2019) 4 [arXiv:1807.08763] [INSPIRE].

[47] G. Louppe, M. Kagan and K. Cranmer, *Learning to pivot with adversarial networks*, arXiv:1611.01046 [INSPIRE].

[48] T. Heimel, G. Kasieczka, T. Plehn and J.M. Thompson, *QCD or what?*, *SciPost Phys.* **6** (2019) 030 [arXiv:1808.08979] [INSPIRE].

[49] K. Kondo, *Dynamical likelihood method for reconstruction of events with missing momentum. 1: method and toy models*, *J. Phys. Soc. Jap.* **57** (1988) 4126 [INSPIRE].

[50] D0 collaboration, *A precision measurement of the mass of the top quark*, *Nature* **429** (2004) 638 [hep-ex/0406031] [INSPIRE].

[51] CDF collaboration, *Measurement of the top quark mass with the dynamical likelihood method using lepton plus jets events with b-tags in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Rev.* **D 73** (2006) 092002 [hep-ex/0512009] [INSPIRE].

[52] P. Artoisenet, V. Lemaitre, F. Maltoni and O. Mattelaer, *Automation of the matrix element reweighting method*, *JHEP* **12** (2010) 068 [arXiv:1007.3300] [INSPIRE].

[53] T. Martini and P. Uwer, *Extending the matrix element method beyond the Born approximation: calculating event weights at next-to-leading order accuracy*, *JHEP* **09** (2015) 083 [arXiv:1506.08798] [INSPIRE].

[54] D.E. Soper and M. Spannowsky, *Finding physics signals with shower deconstruction*, *Phys. Rev.* **D 84** (2011) 074002 [arXiv:1102.3480] [INSPIRE].

[55] D.E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev.* **D 87** (2013) 054012 [arXiv:1211.3140] [INSPIRE].

[56] D.E. Soper and M. Spannowsky, *Finding physics signals with event deconstruction*, *Phys. Rev.* **D 89** (2014) 094005 [arXiv:1402.1189] [INSPIRE].

[57] C. Englert, O. Mattelaer and M. Spannowsky, *Measuring the Higgs-bottom coupling in weak boson fusion*, *Phys. Lett.* **B 756** (2016) 103 [arXiv:1512.03429] [INSPIRE].

[58] D.E. Ferreira de Lima, O. Mattelaer and M. Spannowsky, *Searching for processes with invisible particles using a matrix element-based method*, *Phys. Lett.* **B 787** (2018) 100 [arXiv:1712.03266] [INSPIRE].

[59] S. Prestel and M. Spannowsky, *HYTREES: combining matrix elements and parton shower for hypothesis testing*, *Eur. Phys. J.* **C 79** (2019) 546 [arXiv:1901.11035] [INSPIRE].

[60] B. Kiran, D. Mathew Thomas and R. Parakkal, *An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos*, *J. Imaging* 4 (2018) [arXiv:1801.03149].

[61] D.P. Kingma and M. Welling, *Auto-encoding variational Bayes*, arXiv:1312.6114 [INSPIRE].

[62] P. Vincent, H. Larochelle, Y. Bengio and P.A. Manzagol, *Extracting and composing robust features with denoising autoencoders*, in the proceedings of the 25th *International Conference on Machine Learning (ICML'08)*, July 5–9, New York, U.S.A. (2008).

JHEP10(2019)047

[63] S. Otten et al., *Event generation and statistical sampling for physics with deep generative models and a density information buffer*, arXiv:1901.00875 [INSPIRE].

[64] M. Farina, Y. Nakai and D. Shih, *Searching for new physics with deep autoencoders*, arXiv:1808.08992 [INSPIRE].

[65] J. Hajer, Y.-Y. Li, T. Liu and H. Wang, *Novelty detection meets collider physics*, arXiv:1807.10261 [INSPIRE].

[66] T.S. Roy and A.H. Vijay, *A robust anomaly finder based on autoencoder*, arXiv:1903.02032 [INSPIRE].

[67] K. Joshi, A.D. Pilkington and M. Spannowsky, *The dependency of boosted tagging algorithms on the event colour structure*, *Phys. Rev.* **D 86** (2012) 114016 [arXiv:1207.6066] [INSPIRE].

[68] CMS collaboration, *Search for anomalous $t\bar{t}$ production in the highly-boosted all-hadronic final state*, *JHEP* **09** (2012) 029 [*Erratum ibid.* **03** (2014) 132] [arXiv:1204.2488] [INSPIRE].

[69] ATLAS collaboration, *Search for heavy particles decaying into top-quark pairs using lepton-plus-jets events in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Eur. Phys. J.* **C 78** (2018) 565 [arXiv:1804.10823] [INSPIRE].

[70] ATLAS collaboration, *Search for heavy particles decaying into a top-quark pair in the fully hadronic final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev.* **D 99** (2019) 092004 [arXiv:1902.10077] [INSPIRE].

[71] ATLAS collaboration, *Search for heavy higgs bosons A/H decaying to a top quark pair in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Phys. Rev. Lett.* **119** (2017) 191803 [arXiv:1707.06025] [INSPIRE].

[72] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[73] G. Altarelli, B. Mele and M. Ruiz-Altaba, *Searching for new heavy vector bosons in $p\bar{p}$ colliders*, *Z. Phys.* **C 45** (1989) 109 [*Erratum ibid.* **C 47** (1990) 676] [INSPIRE].

[74] T. Plehn and M. Spannowsky, *Top tagging*, *J. Phys.* **G 39** (2012) 083001 [arXiv:1112.4441] [INSPIRE].

[75] T. Plehn, M. Spannowsky and M. Takeuchi, *How to improve top tagging*, *Phys. Rev.* **D 85** (2012) 034029 [arXiv:1111.5034] [INSPIRE].

[76] Y.L. Dokshitzer, G.D. Leder, S. Moretti and B.R. Webber, *Better jet clustering algorithms*, *JHEP* **08** (1997) 001 [hep-ph/9707323] [INSPIRE].

[77] M. Cacciari, G.P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J.* **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[78] A. Buckley et al., *Rivet user manual*, *Comput. Phys. Commun.* **184** (2013) 2803 [arXiv:1003.0694] [INSPIRE].

[79] ATLAS collaboration, *Data-driven determination of the energy scale and resolution of jets reconstructed in the ATLAS calorimeters using dijet and multijet events at $\sqrt{s} = 8$ TeV*, ATLAS-CONF-2015-017 (2015).

[80] ATLAS collaboration, *Performance of missing transverse momentum reconstruction in proton-proton collisions at 7 TeV with ATLAS*, *Eur. Phys. J.* **C 72** (2012) 1844 [arXiv:1108.5602] [INSPIRE].

JHEP10(2019)047

[81] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, arXiv:1412.6980 [INSPIRE].

[82] F. Chollet et al., *Keras*, https://github.com/fchollet/keras (2015).

[83] M. Abadi et al., *TensorFlow: large-scale machine learning on heterogeneous distributed systems*, arXiv:1603.04467 [INSPIRE].

[84] R. Frederix and F. Maltoni, *Top pair invariant mass distribution: a window on new physics*, *JHEP* **01** (2009) 047 [arXiv:0712.2355] [INSPIRE].

JHEP10(2019)047